Crash Analysis: Avoiding Impact with Insight

By Betty Liu

Capstone Experience in Data Science  
DATA 205 CRN: 32118

**Project overview**   
For this project, the **data** I will be using comes from dataMontgomery. It will be composed of two datasets:   
*Crashing Reporting- Incidents Data:* <https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Incidents-Data/bhju-22kf/about_data> (37 features 109k samples)

*Crashing Reporting- Driver data*: <https://data.montgomerycountymd.gov/Public-Safety/Crash-Reporting-Drivers-Data/mmzv-x632/about_data> (39 features 193k samples)

Both datasets were collected via the Automated Crash Reporting System (ACRS) maintained by the Maryland State Police and include reports filed by the Montgomery County Police, Gaithersburg Police, Rockville Police, and the Maryland-National Capital Park Police.

Using an Outer join, I have combined the two datasets into one big dataset that has 62 features and 217,809 Samples. This data covers reported incidents from January 1st 2015 to April 25th 2025. The dataset is updated weekly.

Features used for the final product of this project include:

Local Case Number, ACRS Report Type, Crash Date/Time, Road Name, Cross-Street Name, Weather, Light, Person ID, Injury Severity, Vehicle Damage Extent, Speed Limit, Latitude, Longitude, Geolocation, Hit/Run.

The most other features were used to during EDA but did not make it to the final product:

Report Number , Agency Name, Route Type, Off-Road Description, Municipality, Related Non-Motorist, Collision Type, Surface Condition, Traffic Control, Driver Substance Abuse, Non-Motorist Substance Abuse, Driver At Fault, Circumstance, Driver Distracted By, Driver’s License State, Vehicle ID, Vehicle First Impact Location, Vehicle Body Type, Vehicle Movement, Vehicle Going Dir, Driverless Vehicle, Parked Vehicle, Vehicle Year, Vehicle Make, Vehicle Model, Lane Direction, Lane Type, Number of Lanes, Direction, Distance, Distance Unit, Road Grade, At Fault, First Harmful Event, Second Harmful Event, Junction, Intersection Type, Road Alignment, Road Condition, Road Division, Season, Year, Month, Time, Holiday, Day Type, Day of Week ,Time of Day.

Some data was also manually collected and entered into spreadsheets, including traffic volume, weather conditions, and road coordinates. This data is used to help normalize the findings when necessary.

The **goal** of this project is to Identify accident hotspots, uncover trends, and analyze risk factors to equip the people of **Montgomery County** and those passing through with information they can use to make safer decisions, and ultimately help create safer roads for everyone. This is significant because the dataset include total of 110,900 reported crashes, and involving approximately 195,692 individuals. These crashes are not just statistics; each data point represents a real person, a father, mother, son, or daughter. The consequences of these incidents are substantial, ranging from personal injuries and vehicle damage to widespread traffic disruptions. In just a matter of minutes, a single crash can impact dozens of lives and affect daily routines across a community. By analyzing this data, we gain valuable insights into driver behavior and environmental factors that influence crash risks also encourage each of us to reflect on and adjust our own behaviors behind the wheel.

The primary **tools** used in this project included Python within Google Colab for all data cleaning and basic exploratory data analysis, Excel for spot-checking values and verifying data integrity, and Tableau Public for creating most of the visualizations. GitHub was used for version control and code documentation. Additionally, I used OpenAI’s ChatGPT for brainstorms, refine technical writing for clarity, and assist in identifying effective search strategies when addressing specific analytical challenges.

**Summary of data cleaning**

The data preprocessing process involved several key steps to prepare the datasets for analysis. First, individual data frames were copied and then merged to create a new dataset. New features were engineered, including time of day, season, holiday status, day type (weekday/weekend), and day of the week, to enhance analysis of temporal patterns. Some feature values were standardized for consistency such as, Agency Name, weather, surface condition, light, traffic control, circumstances, and much more. Missing values were addressed by filling them with "unknown" to retain data completeness.

**Basic descriptive statistics**

The dataset includes 110,900 reported traffic crashes involving approximately 195,692 individuals. Key variables captured in the dataset include crash date and time, location coordinates, weather conditions, lighting conditions, extent of vehicle damage, and crash severity. Rear-end collisions are the most common crash type, accounting for roughly 31% of all incidents. A majority of crashes occurred during weekday rush hours.

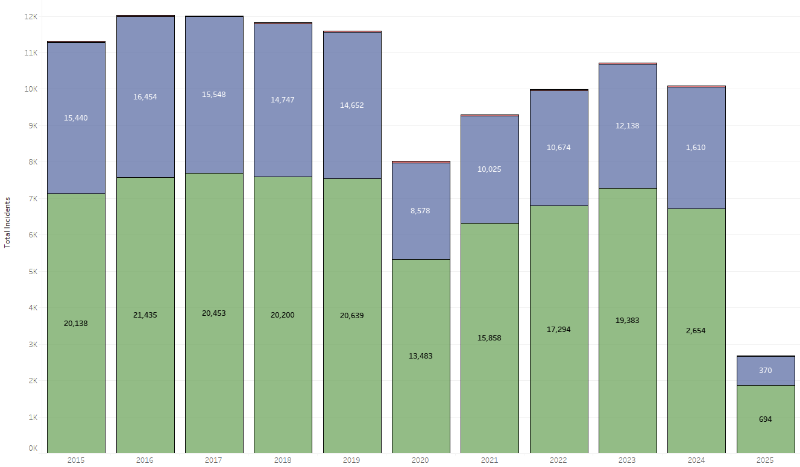
Approximately 68% of crashes took place under clear weather conditions, 12% during rain, and fewer than 2% in snowy conditions. It’s important to note that clear weather does not necessarily indicate greater danger; rather, it reflects the higher frequency of clear days compared to hazardous weather. Additionally, during severe weather, people often reduce travel or stay off the roads entirely, leading to fewer overall incidents in those conditions.

**Description of final data product**

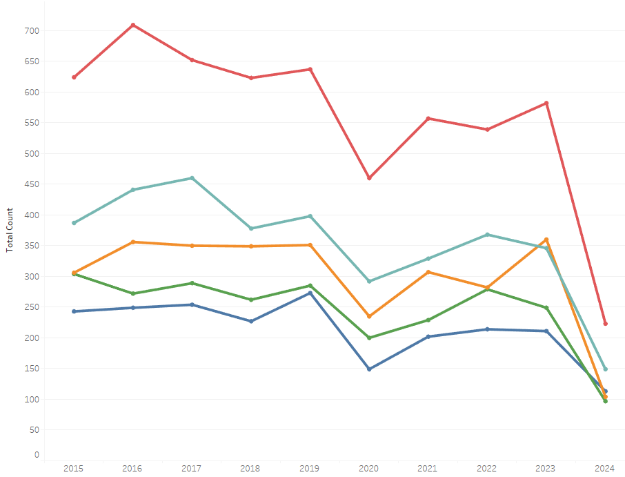
The final data product consists of a comprehensive analysis of traffic crash data in Montgomery County, supported by a series of visualizations.

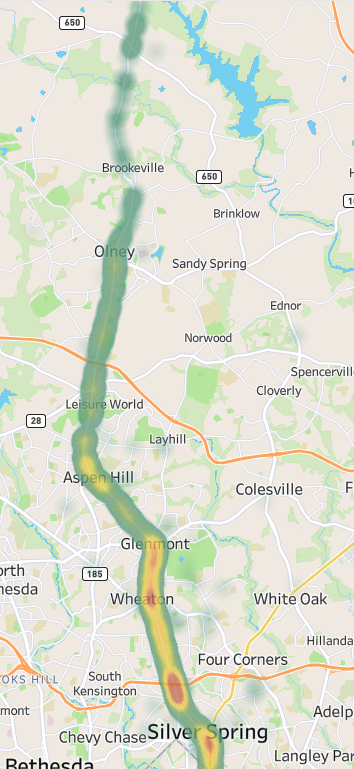
Despite a **decrease** from 2019 of reported incidents by approximately **37.4%**

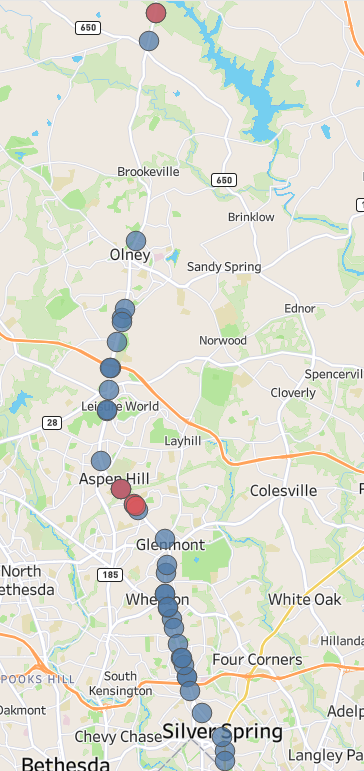
The number of **fatal** incidents **increased** approximately **46.6%**

On the ***left***. Historically, property damage was the most common accident type all being over 60% of the time. There is a clear drop in reported incidents in 2020 and that is likely do to the covid-19 shut downs.

Despite a **decrease** from 2019 of reported incidents by approximately **31%** The number of fatal incidents increased approximately **46.6%.** This is likely due to fewer people being on the roads, which may give some individuals a false sense of confidence and lead them to engage in more reckless behavior. A statistical analysis confirmed that this increase is significant: a two-proportion z-test comparing fatal crash rates between 2019 and 2020 yielded a p-value of 0.0215, indicating that the rise in fatal incidents is unlikely to be due to random chance.

On the ***right***. Historically, the top five roads with the highest number of reported crashes in Montgomery County include Georgia Avenue, New Hampshire Avenue, Frederick Avenue, Rockville Pike, and Connecticut Avenue. Among them, Georgia Avenue has consistently recorded the most crashes. Georgia Avenue is approximately 16.3 miles long, which is about 20 percent shorter than New Hampshire Avenue at 20.5 miles, yet it averages roughly 58% more crashes in the recent years (2020-2024).

The density map on the ***left*** displays crash data for Georgia Avenue from 2020 to 2024. Over these years, the street has typically seen between 28 and 36 crashes per mile, with the exception of 2024, which shows a noticeable drop to around 14 crashes per mile. The highest concentration of crashes is observed in the segment between Aspen Hill and Silver Spring; this area is characterized by higher commercial activity. In contrast, northern stretches of Georgia Avenue are more suburban or wooded, with fewer businesses and generally lower crash density.

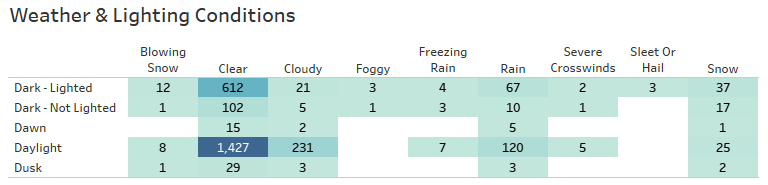


The bubble map on the ***right*** highlights locations where serious injuries were suspected or fatal crashes occurred. One small fatal cluster appears between the Aspen Hill and Glenmont areas, while another isolated incident is located farther out in the Sunshine area. The crash in Sunshine is particularly interesting because it occurred on a stretch of road surrounded only by woods, with no nearby houses or businesses. This section of Georgia Avenue is a two-lane road, with one lane in each direction. The crash took place during daylight hours on a summer evening and involved a head-on collision with two individuals involved in the incident. Given the calm and open conditions, it’s possible the drivers had a false sense of security and were not as engaged in active safe driving.

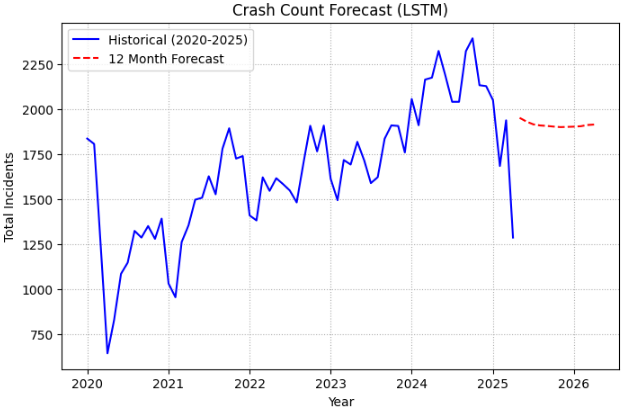
***Below.*** The 30–40 mph range shows the highest concentration of crashes and is also associated with the most severe vehicle damage, including disabling and destroyed vehicles. This suggests that even moderate speed zones can lead to serious consequences in the event of a collision. It's important to note that the “Speed Limit” field in the dataset appears to be somewhat inconsistent- while values above 15 mph generally reflect posted speed limits, lower numbers may represent estimated vehicle speeds at the time of the crash rather than actual speed limits.

It's interesting to note that even at 0 mph, a vehicle can still be severely damaged or disabled. This usually reflects a one-sided crash where the vehicle is either stopped or moving very slowly, such as at 1 or 2 mph, and is struck by another driver. These situations are a strong reminder that no matter how carefully you drive, your safety also depends on the behavior of others. You are only as safe as the least cautious driver around you.

***Below***. Looking at recent data from 2025 and corresponding weather conditions, there were 2,212 crashes over 72 clear days, averaging about 30.7 crashes per clear day. In comparison, 121 crashes occurred over 4 wintry days (~30.2 crashes per day), and 210 crashes were recorded over 11 rainy days (~19.1 crashes per day). At first glance, clear and wintry days appear to pose a similar crash risk.

However, it's important to consider how driver behavior changes during severe weather. On wintry days, many people choose to stay off the roads unless necessary. If the volume of traffic on snowy days were equivalent to that on clear days, we would likely see a much higher number of crashes due to the increased mechanical and environmental challenges of driving in snowy conditions.

To forecast crash trends for the next 12 months, I applied a Long Short-Term Memory (LSTM) model, which is well-suited for time series prediction. The blue line in the graph represents historical crash data, while the red dashed line shows the model's forecast.

***To the right***. The model picks up on patterns in the data over time and suggests that crash numbers might level out after a recent drop. That drop could be caused by unusual events or missing data, but the forecast shows things returning to a more normal range instead of continuing to fall or suddenly rising again. This forecast is helpful for getting an idea of what might happen next, but it’s important not to rely on it too much. Real-world changes and new data could affect what actually happens.

**Vision Zero**. Montgomery County has committed to the Vision Zero initiative, aiming to eliminate all roadway-related fatalities and serious injuries by the year 2030. The county has developed a comprehensive 2030 Action Plan, which outlines specific strategies and actions focused on areas such as road design improvements, speed management, and enhancing pedestrian and bicycle infrastructure. The plan emphasizes a data-driven approach to identify high-risk areas and implement targeted interventions.

Learn More Here: <https://www.montgomerycountymd.gov/visionzero/>

**Conclusion**

This analysis offers a detailed view of traffic crash patterns in Montgomery County. However, the value of this information depends on drivers being mindful of their own behavior and choosing to drive attentively rather than passively. Road safety is a shared responsibility, and we are only as safe as the least cautious driver. By staying engaged and intentional behind the wheel, we can all contribute to shaping a safer future.

**Acknowledgements**

Data Science Faculty  
Professor **Celia Evans** – INTRODUCTION TO DATA SCIENCE – DATA 101.  
Professor **Rachel Saidi** – DATA VISUALIZATION AND COMMUNICATION – DATA 110.  
Professor **Abdirisak Mohamed** – STATISTICAL METHODS IN DATA SCIENCE – DATA 201.  
Professor **Jane Valentine** – CAPSTONE EXPERIENCE IN DATA SCIENCE – DATA 205.

Other MC Staff  
Professor **Gary Thai –** PYTHON PROGRAMMING – CMSC 206.

Special Thanks

**Kathy Luh.** Acting Manager/Data Innovation Team dataMontgomery. (Speed Limit clarification)  
Classmates of DATA 205 in Spring 2025.

**References**

<https://www.gsa.gov/buy-through-us/products-and-services/transportation-and-logistics-services/fleet-management/vehicle-leasing/accidents-and-maintenance/accident-management-center> (Steps after an accident)

<https://maps.roads.maryland.gov/itms_public/?stationid=S2017150112> (Volume Detail Report)

<https://weather.com/weather/monthly/l/4433c656f801f190919b8d258b7dd18125e7b21c99dc22683e1d53839e18ac82> (Weather Details)

<https://www.montgomerycountymd.gov/visionzero/background.html> (Vision Zero Time Line)

<https://www.google.com/maps/place/Montgomery+County,+MD/@39.1361654,-77.3282876,54425m/data=!3m1!1e3!4m6!3m5!1s0x89b6323583b8a387:0x780b190677a96873!8m2!3d39.1547426!4d-77.2405153!16zL20vMGJ4OXk?entry=ttu&g_ep=EgoyMDI1MDUwMy4wIKXMDSoASAFQAw%3D%3D> (map of Montgomery County boundaries)